图书情报工作
第65卷 第3期 2021年2月

基于深度学习算法的学术查询意图分类器构建*

■ 王瑞雪[1] 方婧[1] 桂思思[2] 陆伟[1,3] 张显[4]

[1] 武汉大学信息管理学院 武汉 430072 [2] 南京农业大学信息管理系 南京 210095

[3] 武汉大学信息检索与知识挖掘研究所 武汉 430072 [4]百度时代网络技术(北京)有限公司 北京 100085

摘 要:[目的/意义]实现学术查询意图的自动识别,提高学术搜索引擎的效率。[方法/过程]结合已有查询意图特征和学术搜索特点,从基本信息、特定关键词、实体和出现频率4个层面对查询表达式进行特征构造,运用Naive Bayes、Logistic回归、SVM、Random Forest四种分类算法进行查询意图自动识别的预实验,计算不同方法的准确率、召回率和F值。提出了一种将Logistic回归算法所预测的识别结果扩展到大规模数据集、提取"关键词类"特征的方法构建学术查询意图识别的深度学习两层分类器。[结果/结论]两层分类器的宏平均F1值为0.651,优于其他算法,能够有效平衡不同学术查询意图的类别准确率与召回率效果。两层分类器在学术探索类的效果最好,F1值为0.783。

关键词:学术查询意图 自动识别 两层分类器

分类号:G250.2

DOI:10.13266/j.issn.0252-3116.2021.03.012

1 引言

随着科学文献等学术资源的爆炸增长[1],为快速获取学术信息、方便学习生活、提高科研工作效率,专业学术搜索引擎从面对少量专业用户发展为面向大量的、不同类型的用户;从面对单一的科研需求发展为面向复杂的多样性需求[2]。与此同时,由于学术数据库(Web of Science、CNKI等)的检索系统的专业化,各数据库间不能共享互联,"一站式"学术搜索引擎如百度学术、谷歌学术成为学术查询的首选[3-4]。由于用户的学术背景和学术能力不同,在进行学术查询时其需求往往不同。对于使用学术搜索引擎的科研用户,其需要多样化的学术信息,了解研究进展、追踪研究前沿,简短的查询表达式并不足以准确表达其学术查询意图;对于学术新手或非科研用户使用学术搜索引擎时,因其对相关学术领域的了解较浅,进行学术查询时的关键词并不准确,从而不能获取精准的学术信息。

学术查询意图为用户进行学术搜索时通过查询表达式所表达的用户信息需求。现有的学术搜索引擎多为基于关键词匹配的搜索技术,不能识别学术查询语句的查询意图,而通过对学术查询意图的识别,可为进一步优化学术搜索结果、提高检索效率、节约用户时间,使得用户的学术搜索体验得以提高、获取更精准的学术信息。

查询意图自动识别多采用文本分类的算法,研究人员通常针对一个分类算法,通过组合不同的特征,试验查询意图识别的效果。实验的基础是查询意图的类目体系,而针对学术查询意图的分类大多在普通查询意图的基础上展开,缺乏学术查询意图的专门类目体系,在此基础上,本文的前序基础研究构建了学术查询意图类目体系,共分为5类:

(1)学术文献类:指用户通过查询式获取某篇特定的学术出版文献,如通过学术文献的标题构造查询式进行查询。

(2)学术实体类:指用户通过查询获取学术实体的相关信息,如科研机构实体"山东省农业科学院作物研究所"。

(3)学术探索类:指用户在某个领域进行探索性查询,需多次交互查询才可获取所需的学术资源,如对学术概念词"神经网络算法"进行查询等。

* 本文系国家社会科学基金青年项目"面向学术搜索的查询意图研究"(项目编号:19CTQ023)研究成果之一。

作者简介:王瑞雪(ORCID:0000-0001-5932-9036),博士研究生,E-mail:ruixue_wang@whu.edu.cn;方婧(ORCID:0000-0002-9538-7812),硕士研究生;桂思思(ORCID:0000-0001-7562-7447),讲师,博士;陆伟(ORCID:0000-0002-0929-7416),院长,教授,博士;张显(ORCID:0000-0002-8274-9523),硕士研究生。

收稿日期:2020-06-17 修回日期:2020-10-14 本文起止页码:93-99 本文责任编辑:杜杏叶

（4）知识问答类：指用户通过查询获取某个特定问题的答案，如“亚急性甲状腺炎治愈后会复发吗？”等问题。

（5）非学术文献类：指用户通过查询希望获取一些政策、行业报告等非学术文献，如查询“汽车下乡的政策”等。

为了适度提高学术查询意图分类的召回率，本研究构建了基于 Logistic 回归算法的两层分类器来实现学术查询意图的自动识别，提出了适用于学术查询意图识别的 4 个特征：基本信息、词中信息、实体信息和出现频率，并在百度学术查询日志数据集进行测试和评价。

2 相关研究

关于学术查询意图识别的研究较少，多为面向图书查询的意图识别研究或应用，如胡伶霞[5]将图书检索的查询意图分为单意图与多意图并利用词典对检索词进行分类，李兵[6]借鉴与完善胡伶霞[5]的图书检索的查询意图体系用以提高图书分面检索的效果。针对学术查询意图识别研究较少且多聚焦在图书查询意图研究这一现状，考虑到学术搜索是搜索的一个特定垂直搜索子领域，因此可借鉴综合搜索引擎中的用户查询意图识别研究。

以综合搜索引擎中的用户查询意图为研究对象的查询意图识别分类方法可分为人工识别方法[7-8]和自动识别方法[9]。由于查询意图人工识别成本较大，多数研究采用自动识别方法，涉及的分类特征以及分类算法归纳如下：

2.1 查询意图自动识别的分类特征来源

查询意图自动识别的分类特征来源包括“搜索引擎检索结果的点击行为、用户查询表达式”等。J. Brenes 等[10]指出用户的点击行为是查询意图分类的最有效特征，但点击行为数据涉及用户隐私，存在获取权限的问题，仅有 Y. Liu 等[11]的实验中利用了点击分布作为特征进行查询意图识别。研究人员大多通过对查询表达式进行分析以获取综合搜索中查询意图的分类特征，可归纳概括为“基本信息、词中信息和实体信息”三类。

（1）基本信息：指查询表达式的长度、词项个数、词项长度等基本信息，比如 N. Belkin[12]通过实验得出查询表达式词长为 2 以下的意图大概率为导航类，词长越长的查询表达式为信息类查询的概率越大。

（2）词中信息：指查询表达式中所含的词汇信息。研究者们对该类特征的研究较多，比如 B. Jansen[13]和 M. Herrera[14]先后总结了英文综合搜索中关键词特征与查询意图的对应情况；张晓娟[15]总结了中文综合搜索查询中不同意图类别查询的特征词。针对学术搜索，M. Khabsa 等[16]通过对学术搜索引擎 CiteseerX 的查询意图分析，总结了学术搜索中导航类查询的特征词，例如是否包含年份、是否包含标点符号、是否包含停用词等。尽管利用查询表达式中的词汇信息作为特征的方法比较简单，但众多研究的结果表明该类特征对查询意图的识别较为有效。

（3）实体信息：指查询表达式中所含有的实体信息，比如张晓娟[15]将实体与查询意图进行分类对应，总结出导航类查询中的实体多为人名、地名、机构名，资源类查询中实体多为游戏名、歌曲名等；Y. Chang 等[17]将实体的自然语言处理结果作为分类的特征。

2.2 查询意图自动识别算法

查询意图自动识别算法采用基于查询表达式特征的方法对查询意图进行识别，其本质上是一种文本分类方法，在进行分类器选择时，研究者往往会根据实验数据集、实验数据特征和具体分类任务的情况选择不同的分类算法，如 SVM、决策树、PLAS 等，例如，Y. Liu 等[11]使用典型决策树算法将 nCS、nRS 和点击分布三种特征结合起来执行识别任务；M. Mendoza[18]利用 SVM 与 PLSA 对查询意图进行归类；Y. Chang 等[17]提出了使用自然语言处理（NLP）的分析结果作为特征进行查询意图分类的方法，取得了较好的结果。查询意图自动识别的特征与方法大多针对综合搜索引擎，较少关注学术搜索这一垂直细分领域，由于学术搜索的专业性与特殊性，其特征与方法并不能直接适用于学术查询意图的自动识别，需在查询意图识别的基础上进一步扩展，以适应学术查询意图的自动识别。

3 特征选择与分类器构造

3.1 学术查询意图的特征提取

综合搜索中查询意图的特征可从查询词中获得[13]，可分为基本信息[12]、词中信息[13-16]和实体信息[15,17]三类，结合学术搜索的特点，可将学术查询意图的分类特征扩展为基本信息、词中信息、实体信息和词汇出现频率的统计特征 4 个方面。由于百度学术中约占 30% 的查询表达式为英文，因此学术查询意图分类特征时对中英两种语言进行了综合考量。

3.1.1 基本信息特征

基本信息特征是指可从查询表达式中直接提取的

王瑞雪，方婧，桂思思，等.基于深度学习算法的学术查询意图分类器构建[J].图书情报工作,2021,65(3):93－99.

信息,如查询表达式的长度、词项个数、词项长度,该特征可分为两个维度:字符特征与词项特征。

字符特征包括:查询表达式字符数和字符占比。前者是指查询表达式中不同类别字符的绝对数值;后者是指不同类别字符在查询表达式的相对比例。查询表达式字符分为中文字符、英文字符、标点符号字符、其他字符(除中/英/标点之外的字符)四类,在统计绝对数值时,每个字符都计数为1。

词项特征包括:查询表达式词项数和查询表达式占比。前者是指查询表达式中不同类别词项的绝对数值;后者是指不同类别词项在查询表达式的相对比例。由于中英文的区别,在实验过程中,对中文查询表达式进行分词的预处理,按分词结果计算中文的词项数量;同时将单个英文单词算为一个词项。

查询表达式的基本信息维度四类特征如表1所示:

表1 基本信息维度特征

类别	特征
查询表达式字符数	总字符数;中文字符数;英文字符数;标点符号字符数;其他字符数
查询表达式字符占比	中文字符占比;英文字符占比;标点符号占比;其他字符占比
查询表达式词项数	总词项数;中文词项数;英文词项数;其他词项数
查询表达式词项占比	中文词项占比;英文词项占比;其他词项占比

3.1.2 词中信息

查询表达式所含特定词常被作为识别用户查询意图的重要特征[16]。本研究根据“百度学术”用户查询日志情况,发现学术文献类和知识问答类有较为明显的词中信息特征。例如,对于学术文献类,发现查询表达式中经常直接使用包含引文格式的相关信息(如:使用[J]、[C]等中文引文规范来表示杂志和会议文献);对于知识问答类,发现其查询表达式中疑问词使用较多;具体如表2所示:

表2 词中信息维度特征

类别	特征
学术文献类:参考文献著录特征	含有年份(如:2005);含有书名号/双引号;含有中文引文相关(如:包含[J]/[C]/[M]/[D]);含有英文引文相关(如:包含 et al./ACM/Springer/Emerald/Elsevier/ Press 等)
知识问答类:疑问词特征	含有中文疑问词(如:谁、什么、哪里、几时、多少、怎么、为什么、是否、能否等);含有英文疑问词(如:who, what, which, whose, when how, why 等);含有其他疑问相关词(如:试析、浅论、原因、区分等)

3.1.3 实体特征

在综合搜索的查询式中经常会出现命名实体,J. Guo[19]研究发现英文综合搜索中有70%的查询表达式中包含命名实体。同时由于本文的前序研究的学术查询意图类目体系中有学术实体类,因此本研究将实体信息作为一个特征,主要用于识别学术实体类的查询表达式。具体来说,作为特征的命名实体包括以下四种:人名、地名、机构名、学术实体(如:杂志、大学、研究所、研究院、中心、实验室等)。

3.1.4 词汇出现频率的统计特征

在本研究的人工标注实验过程中发现学术专有名词(例如,氨基酸、苯)在学术探索类查询表达式中出现的频率较高,而日常词汇在知识问答类查询式中的词语出现频率较高。为了描述该类特征,本研究借鉴 Inverse Document Frequency(IDF)的概念,提出了一个出现频率S(W)的指标,该指标对词语的出现频率特征进行量化。对于任何一个词语W,

$$S(W) = \log\left(\frac{n+1}{N(W)+1}\right) \qquad 公式(1)$$

其中n是数据集中查询表达式Q的总数量;N(W)表示词语W出现在数据集的查询表达式的频次,取值范围为[1,n]。如果一个词语出现在所有的查询表达式中,那么其出现频率S(W)的值为0;如果一个词语仅仅出现在一个查询表达式中,那么其出现频率S(W)的值为最大值 $\log\left(\frac{n+1}{2}\right)$。

对数据集中的所有查询表达式中的每个词语W计算其出现频率S(W),针对每个查询表达式Q计算出每个查询表达式的最大出现频率 $S_{MAX}(Q)$、最小出现频率 $S_{MIN}(Q)$ 和平均出现频率 $S_{AVE}(Q)$ 三个特征,计算公式为:

$$S_{MAX}(Q) = \max_{W \in Q} S(W) \qquad 公式(2)$$

$$S_{MIN}(Q) = \min_{W \in Q} S(W) \qquad 公式(3)$$

$$S_{AVE}(Q) = \frac{\sum_{W \in Q} S(W)}{Count(W)}$$,Count(W)表示统计Q中不同W的数量。

3.2 二层分类器的构建

针对实验数据的训练集较小的特点,为提高结果准确度,本研究对于五类学术查询意图中的每个类别采用二元分类法:针对于每个学术查询意图类别,本研究构建了一个监督学习算法,将查询表达式表示成一系列的特征X,在目标空间中寻找到一个最优化函数F可根据特征X预测查询表达式是否属于的查询意图类别y。该过程也可以简单的表示成以下公式:

$$y = F(x) \qquad 公式(4)$$

查询表达式的特征抽取及训练集规模将会影响监督学习算法的效果。由于本研究可用于监督学习算法训练的数据集较小,为提升效果将从以下两个方面改善:①抽取可以表征训练数据的合适特征,②获取足够多的训练数据集。基于此,本研究提出一种两层分类器来实现学术查询意图的自动识别,如图 1 所示:

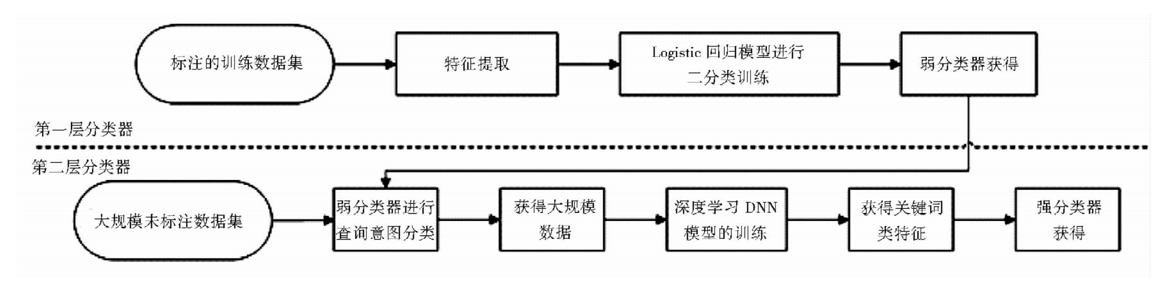


图 1　两层分类器算法逻辑示意

在第一层,实验人员从 4 000 条的标注数据中,采用 Logistic 回归算法根据查询表达式的四类数据特征训练出一个分类器。由于分类结果的召回率较低,称该阶段的分类器称为弱分类器。采用二项 Logistic 回归模型对每一个输入的查询表达式依次做二元判断是否属于学术文献类、学术实体类、学术探索类、知识问答类、非学术文献类这五个查询类别,在标注数据集上进行分类器训练,得到弱分类器。

第二层为深度学习分类器,采用深度学习的 DNN 模型进行训练与分类。首先,利用第一层的弱分类器对大规模未标注数据进行分类标注,对查询表达式的学术查询意图类别进行预分类;其次,预分类的标注数据作为第二层分类器的输入数据,用于 DNN 模型的训练与分类。在大规模数据集下,第二层分类器仍能够学习到第一层弱分类器中并未提及的数据特征,将这部分在大规模数据集下学习到的新特征统称为“关键词类”特征,用于提高本层分类器的效果。例如,有两个查询表达式“关于春天的谚语有哪些”“关于秋天的谚语有哪些”,由于都含有词中信息类特征“哪些”,弱分类器将会把他们归类为知识问答类。伴随大规模数据集中这类查询表达式出现频次的提高,深度学习分类器将会学习到“关于 XX 的谚语”是重要的关键词,当遇到“关于龙的谚语”这类查询查询式时,深度学习分类器依据新提取的“关键词类”特征,会将其归为知识问答类。本实验采用 Python 代码构建,调用 TensorFlow 开源软件库。

4　实验

4.1　实验数据

本研究获得了由百度学术提供的查询日志为实验数据,该日志记录了用户在“百度学术”搜索栏中所有的交互信息,每一条数据记录了用户的唯一标识符(UID)、查询时间(Time)、查询表达式(Query)以及查询 IP 地址(IP),见图 2。数据总量为 5 414 886 条,剔除乱码数据、重复数据后,数据总量为 3 449 591 条,其中 1 000 条数据已由原论文作者按照前序研究的学术查询意图类目体系标注了相应的类别,本文将这 1 000 条数据作为测试集。

1cc6aace735c0285f62d345fb00d3e4f	2018-03-01 22:13:18	刘洋(2005)	101.94.11.33
bc4c5691deef09e65cd227eab4052572	2018-03-01 00:00:46	卵巢畸胎瘤恶性严重吗	123.147.246.98
151d10ac97f55568e4697b6164384e7d	2018-03-01 13:59:01	车牌识别	180.169.121.82
0b8d163d6c968d58cf196c07e6d429ec	2018-03-01 13:35:50	裂流研究综述	219.230.160.122

图 2　百度查询日志数据格式样例

在清洗后的数据集中,笔者随机抽取与测试集不同的 4 000 条数据作为训练集,招募了情报学专业研究生一年级,且有相关标注工作经验的 6 名同学,将 4 000条学术查询数据标注为“学术文献类”“学术实体类”“学术探索类”“知识问答类”和“非学术文献类”五个类别。具体过程如下:

(1)为标注者介绍了标注任务背景、实验逻辑和其标注的数据集的使用背景。

(2)编写《学术搜索查询意图人工标注指南》,介绍学术查询意图类目体系,以 1 000 条测试集中实例说明了 5 种学术查询意图的界限,使标注者对查询意图的分类有大体感知。

(3)每两人一组,独立根据上述要求,对全部分配的数据进行类别标注。前两组每组分配 1 340 条数

据，第三组分配 1 320 条，合计 4 000 条数据。允许标注者在产生标注疑惑时可借助百度的查询结果页内容进行判断。

完成人工标注任务后，笔者采用 Kappa 系数来衡量标注结果之间的一致性，对上述 3 组的标注结果进行了一致性检验，Kappa 值分别为 0.776、0.759、0.806。Kappa 值均高于 0.75，说明标注者之间分类判别的一致性较高。对于标注结果不同的数据，笔者后续召集了所有的标注者对其进行讨论，并按照多数性原则最终类别。

4 000 条查询表达式的学术查询意图标注结果按查询意图类别的统计分析结果如下表 3 所示：

表 3　训练集学术查询意图分类比例结果统计

类别	学术文献类	学术实体类	学术探索类	知识问答类	非学术文献类	总计
数目	1058	161	1845	561	341	3 966
占比/%	26.68	4.06	46.52	14.14	8.60	100

所标注的 4 000 条数据中，有 34 条数据为完全没有语义的文字或纯标点符号，不构成学术查询表达式，无法将其归类为 5 类学术查询意图，故对其剔除，其余 3 966条数据进行分类统计，可知学术查询意图为“学术探索类”的查询表达式比例最高。

下图 3 是对本次实验人工标注的训练集（3 966 条）和测试集（1 000 条）的学术查询意图类别比例进行对比分析结果。可以看到，两次人工标注的数据集中，“学术文献类、学术实体类、知识问答类”三类的比例基本一致；“学术探索类和非学术文献类”存在 2%左右差距。数据集的不同和标注者对标注说明理解不同都会造成标注结果一定程度的误差，但平均比例误差为 1.13%，说明训练集与测试集中各类型数据所占比例相似，可排除模型误差。

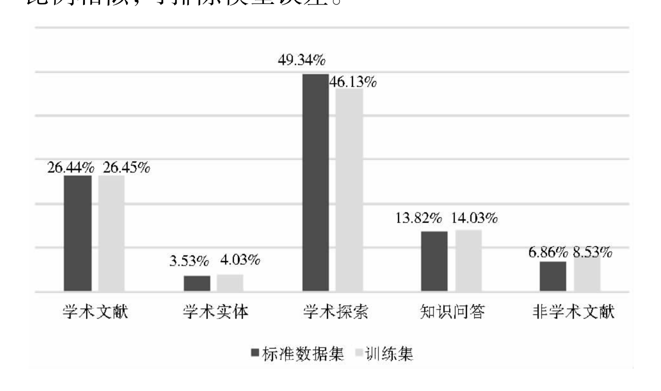


图 3　训练集与测试集中学术查询意图分类比例对比图

4.2　实验结果

表 4 统计了本研究中 4 个预实验和 1 个正式实验不同算法的统计结果。从表中的数据可以看出，在宏平均准确率方面，SVM 效果最佳为 0.789，在宏平均召回率方面，Naive Bayes 效果最佳为 0.755。但同时可以发现，单层分类器并不能很好的兼顾准确率和召回率两个指标，其中大多数算法的召回率值偏低；而本研究所提出的两层分类器在保持准确率的基础上，提升召回率，最终达到了宏平均准确率为 0.767、宏平均召回率为 0.586、宏平均 F1 值为 0.651 的效果。由此说明了，本文所提出的两层分类器效果优于其他 4 种单层分类器算法。

表 4　不同算法的实验效果

分类器	宏平均 P 值	宏平均 R 值	宏平均 F1 值
Naive Bayes	0.389	0.755	0.489
Logistic 回归	0.752	0.524	0.599
SVM	0.789	0.495	0.588
Random Forest	0.702	0.546	0.603
二层分类器	0.767	0.586	0.651

表 5 展示了本研究所提出的二层分类器在各学术查询意图类别的分类情况。由表 5 中数据分析可以发现，二层分类器在学术文献和学术探索两类上分类效果较好，在学术实体类和知识问答类分类效果一般，但仍高于其他单层分类器。

表 5　二层分类器最终实验效果

类别	P 值	R 值	F1
学术文献类	0.758	0.603	0.672
学术实体类	0.727	0.444	0.551
学术探索类	0.738	0.834	0.783
知识问答类	0.845	0.462	0.597
宏平均	0.767	0.586	0.651

从 4 种学术查询意图类别的维度出发，对比分析每个类别运用不同分类器的分类效果。

（1）学术文献类：表 6 呈现了 4 个预实验分类器和 1 个二层分类器在学术文献这一类别的实验结果。从 F1 值上看，采用 Random Forest 算法的单层分类器效果最佳（0.697），本研究提出的二层分类器效果次之（0.672）。总体来说，5 种分类器在学术文献类查询式的自动识别上效果良好。

表 6　不同分类器在学术文献类的效果

分类器	P 值	R 值	F1
Naive Bayes	0.565	0.667	0.612
Logistic 回归	0.748	0.602	0.667
SVM	0.804	0.567	0.665
Random Forest	0.721	0.674	0.697
二层分类器	0.758	0.603	0.672

在已有研究中 Y. Khabsa[17] 等对学术搜索引擎 CiteseerX 进行查询意图分析时,对其中导航类查询意图(占比为 12.5%)做了自动识别研究。在其定义框架下,学术搜索中导航类查询为“用户查询意图为希望搜索到某种出版物”,与本研究的“学术文献类”定义一致。针对该类查询,Y. Khabsa 采用 GBT 算法、以人工标注的 579 条数据为训练集训练分类器,其实验结果为:准确率 0.68、召回率 0.68、F1 值 0.677。本研究的“学术文献类”实验结果与之基本持平,在 F1 指标上略有提升。

(2)学术实体类:表 7 呈现了 4 个预实验分类器和 1 个二层分类器在学术实体这一类别的实验结果。本研究所提出的二层分类器在 P 值和 F1 值上均比预实验分类器效果好,但总体上,学术实体类的查询意图识别在四种学术查询意图类别中的效果最差,有以下两个原因:一是本研究实体识别的工具采用 Stanford 实体识别工具,中文实体识别效果不佳,而中文查询表达占数据总量的 70%;二是本研究在类别分类时规定,仅含有一位学者的查询表达式归为学术实体类,含有多个学者名的查询表达式归为“学术全文类”,导致学术实体类的自动识别不是简单的实体识别问题,还包含判断查询表达式中学者数量的问题,从而导致了该类别学术查询意图识别效果较低。

表 7 不同分类器在学术实体类的分类效果

分类器	P 值	R 值	F1
Naive Bayes	0.247	0.543	0.339
Logistic 回归	0.688	0.314	0.431
SVM	0.714	0.286	0.408
Random Forest	0.632	0.343	0.444
二层分类器	0.727	0.444	0.551

(3)学术探索类:表 8 呈现了 4 个预实验分类器和 1 个二层分类器在学术探索这一类别的实验效果。二层分类器在 F1 值上得分最高,其准确率和召回率得分也较高,说明该分类其在学术探索类识别的效果较好,该类查询表达式占总体查询表达式的比例最大,为 46% 左右。

表 8 不同分类器在学术探索类的分类效果

分类器	P 值	R 值	F1
Naive Bayes	0.524	0.864	0.652
Logistic 回归	0.716	0.767	0.741
SVM	0.749	0.749	0.749
Random Forest	0.711	0.721	0.716
两层分类器	0.738	0.834	0.783

(4)知识问答类:表 9 呈现了 4 个预实验分类器和 1 个二层分类器在知识问答这一类别的实验结果。可以看到,除了 Naive Bayes 方法外的召回率都普遍较低;本文所提出的两层分类器总体上看(F1 值)还是略好于其余方法。

表 9 不同分类器在知识问答类的分类效果

分类器	P 值	R 值	F1
Naive Bayes	0.219	0.946	0.355
Logistic 回归	0.857	0.415	0.559
SVM	0.891	0.377	0.530
Random Forest	0.744	0.446	0.558
两层分类器	0.845	0.462	0.597

5 总结与展望

本研究聚焦学术查询意图的分类研究,通过对学术查询表达式进行分析,基于已有研究对查询表达式从基本信息、词中信息词、实体信息和词汇出现频率的统计特征四个方面进行基础的特征描述,构建了针对学术查询进行查询意图自动识别的两层分类器,并基于大规模数据的分类特征提取了“关键词类”特征。对比其他单层分类器相比,本研究提出的两层分类器在宏平均 F1 值上取得较好结果,能够有效兼顾不同查询意图类别的准确率与召回率。

本研究的不足之处在于,由于针对学术查询意图研究的成果相对较少,缺乏统一的、大规模的评测数据集,因此,本研究的两层分类器效果难以与其他实验结果进行横向对比。下一步将着重推广学术查询意图自动识别的相关数据集,促进不同方法针对学术查询意图的自动识别的横向对比。

参考文献:

[1] BORNMANN L, RÜDIGER M. Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references[J]. Journal of the Association for Information Science and Technology, 2015, 66(11): 2215 – 2222.
[2] 周剑, 王艳, XIE I. 世代特征,信息环境变迁与大学生信息素养教育创新[J]. 中国图书馆学报, 2015, 41(4):25 – 39.
[3] DONG X, GABRILOVICH E, GEREMY H, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion[C]// Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. New York: ACM, 2014:601 – 610.
[4] 赵蓉英, 陈烨. 学术搜索引擎 Google scholar 和 Microsoft academic search 的比较研究[J]. 情报科学, 2014, 32(2):3 – 6,15.
[5] 胡伶霞. 图书馆 OPAC 检索中基于词典的查询意图自动识别[J]. 图书馆学研究, 2016(23):72 – 76.
[6] 李兵. 基于查询意图识别的自适应图书分面检索研究[J]. 图

书馆学研究，2017(15):57－64.
[7] BRODER A. A taxonomy of web search[C]//Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval. Tampere: ACM, 2002: 3－10.
[8] ROSE D, LEVINSON D. Understanding user goals in web search [C]//Proceedings of the 13th international conference on World Wide Web. New York: ACM, 2004: 13－19.
[9] BEITZEL S, JENSEN E, FRIEDER O, et al. Automatic web query classification using labeled and unlabeled training data[C]// Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. Salvador: ACM, 2005:581－582.
[10] BRENES D, GAYO-AVELLO D, PÉREZ-GONZÁLEZ K. Survey and evaluation of query intent detection methods[C]// Proceedings of the 2009 workshop on web search click data. Barcelona: ACM, 2009:1－7.
[11] LIU Y, ZHANG M, RU L, et al. Automatic query type identification based on click through information[C]//Asia information retrieval symposium. Singapore: Springer, 2006: 593－600.
[12] BELKIN N, KELLY D, KIM G, et al. Query length in interactive information retrieval[C]//Proceedings of the 26th annual international ACM SIGIR conference on research and development in information retrieval. Toronto: ACM, 2003: 205－212.
[13] JANSEN B, BOOTH D, SPINK A. Determining the user intent of web search engine queries[C]//Proceedings of the 16th international conference on World Wide Web. Banff: ACM, 2007: 1149－1150.
[14] HERRERA M. R, MOURA E. S, CRISTO M, et al. Exploring features for the automatic identification of user goals in web search [J]. Information processing & management, 2010, 46(2): 131－142.
[15] 张晓娟. 查询意图自动分类与分析[D]. 武汉:武汉大学, 2014.
[16] KHABSA M, WU Z, C. GILES L. Towards better understanding of academic search[C]// Joint conference on digital libraries 2016. Newark: ACM, 2016: 111－114.
[17] CHANG Y, HE K, YU S, et al. Identifying user goals from Web search results [C]//International conference on Web intelligence Hong Kong: ACM, 2006: 1038－1041.
[18] MENDOZA M, ZAMORA J. Identifying the intent of a user query using support vector machines [C]//International symposium on string processing and information Retrieval. Berlin: Springer, 2009: 131－142.
[19] GUO J, XU G, CHENG X, et al. Named entity recognition in query[C]//Proceedings of the 32nd international ACM SIGIR conference on research and development in information retrieval. Boston: ACM, 2009: 267－274.

作者贡献说明：

王瑞雪：实验设计、数据清洗与论文起草；
方婧：数据清洗、实验操作与论文起草；
桂思思：实验设计与论文修订；
陆伟：实验设计与论文修订；
张显：数据清洗与实验设计。

Based on Deep Learning Algorithm to Construct the Classifier of Academic Query Intent

Wang Ruixue[1] Fang Jing[1] Gui Sisi[2] Lu Wei[1,3] Zhang Xian[4]
[1] School of Information Management, Wuhan University, Wuhan 430072
[2] College of Information Science & Technology, Nanjing Agricultural University, Nanjing 210095
[3] Institute for Information Retrieval and Knowledge Mining, Wuhan University, Wuhan 430072
[4] Baidu Times Network Technology (Beijing) Co., Ltd. Beijing 100085

Abstract: [**Purpose/significance**] To find the solutions of automatically identifying search query intent and improve the efficiency of academic search engines. [**Method/process**] Combining the features of query intent and academic search, we constructed the feature from four aspects, which are the basic descriptive statistics, the special keywords, entity information and the frequency. For the experiments, we examined four types of classifiers which are the Naive Bayes, Logistic regression, SVM, Random Forest and calculated precision, recall and F-measure. A method which is extending the recognition results of academic query intent predicted by Logistic regression algorithm to large-scale data sets and extracting "keyword type" features is proposed to construct a two-layer classifier based on deep learning algorithm for academic query intent recognition. [**Result/conclusion**] The macro-average F1 value of the two-layer classifier is 0.651, which is superior to other algorithms. This method can effectively balance the precision and recall rate of different academic query intentions. The final second-layer prediction model receives the best classification performance, the score of F1 is 0.783.

Keywords: academic query intent automatic identification two-layer classification